Title
Computational phylogenetics and the classification of South American languages

Journal
Language and Linguistics Compass, 13(12)

ISSN
1749-818X

Authors
Michael, L
Chousou-Polydouri, N

Publication Date
2019-12-01

DOI
10.1111/lnc3.12358

Peer reviewed

Computational Phylogenetics and the Classification of South American Languages

Abstract

In recent years, South Americanist linguists have embraced computational phylogenetic methods to resolve the numerous outstanding questions about the genealogical relationships among the languages of the continent. We provide a critical review of the methods and language classification results that have accumulated thus far, emphasizing the superiority of character-based methods over distance-based ones, and the importance of developing adequate comparative datasets for producing well-resolved classifications.

Key words: computational phylogenetics, language classification, South American languages

1. Introduction

South America presents one of the greatest challenges to linguists seeking to unravel the genealogical relationships among the world's languages, exhibiting one of the highest rates of linguistic diversity of the world's major regions (Epps 2009, Epps and Michael 2017). The comparative study of South American languages is quite uneven, however, with most language families lacking classifications based on the comparative method, and the internal organization of even major families being uncertain in important respects.

Given the magnitude of the challenge facing South American linguists, it is perhaps unsurprising that they have embraced the promise of computational phylogenetic methods for clarifying genealogical relationships. The goal of this paper is to provide a state-of-the-art overview of this active research area, focusing on the strengths and weaknesses of the various methods employed, and their contributions to the classification of South American languages.

As we show, computational phylogenetic methods are already yielding important results regarding the classification of South American languages, and the prospects for future advances are promising. At the same time, we argue that challenges remain for the productive application of these methods, and that linguists need to be cognizant of the relative strengths and weaknesses of different methods. In particular we discuss the strengths of character-based phylogenetic methods in comparison to distance-based ones, and the importance of developing datasets adequate for resolving the genealogical relations of interest.

We emphasize that the adoption of the computational methods described in this paper in no way renders obsolete traditional approaches to the study of genealogical relationships; the comparative method exhibits important strengths that are complementary to those of computational phylogenetic methods. Extant phylogenetic methods are incapable of yielding morphological and phonological reconstructions typical of comparative-historical research, as the latter critically rely on historical linguists' accumulated knowledge regarding likely trajectories of historical change.

2. Computational Approaches to Genealogical Relationships

Computational phylogenetic (CP) methods fall into two major groups: distance-based methods and character-based ones (see Nichols and Warnow 2008 and Dunn 2015 for brief overviews and Felsenstein 2004 for more details; see Drummond and Bouckaert 2015 for Bayesian Inference.). Distance-based methods yield classifications based on overall similarity between languages, while character-based methods yield genealogical classifications on the basis of shared innovations, and are thus congruent with the traditional comparative method. We briefly describe these two types of methods, compare their strengths, and argue that only character-based methods reliably recover genealogical relationships among languages.

All CP methods are based on the analysis of a set of *characters*, coded features of the taxa (languages, in our case) whose genealogical relationships we seek to determine. Characters are informative about genealogical relationships when they are based on *homologous* features, i.e. features that are shared by taxa via inheritance from a common ancestor (see DeSalle (2006) for a review of character choice and coding). Good examples of homologous features in linguistics include exhibiting members of a particular cognate set (i.e. reflexes of a proto-form), or the members of a sound correspondence set.

All phylogenetic methods require as input a character matrix, i.e., a table of character values for all taxa under consideration. The construction of the character matrix is arguably the most important and time-consuming step of a phylogenetic analysis, involving the selection of characters, the selection of character states, and the coding of the languages for these characters.[1] This first phase of a phylogenetic analysis draws heavily on the linguist's experience, observation, intuitions, and opinions. Most linguistic phylogenetic analyses until now have been based on lexical data (often Swadesh lists), with characters consisting of the presence/absence of a member of a given *root-meaning set* (Chang et al. 2014), where a root-meaning set is a cognate set in which all members of the set share the same meaning.[2] Languages exhibit a vast range of homologous features (phonological, lexical, morphological, etc.), however, all of which are potentially useful for subgrouping purposes.

Once a character matrix has been developed, it can be analyzed using either distance-based or character-based methods. Distance-based methods generally proceed in two steps:

1. A pairwise *distance matrix* is derived from the character matrix by calculating the distance between every pair of languages in the sample, based on a chosen metric. Commonly used distance metrics in linguistics include the percentage of shared cognates and the summed Levenshtein distances between words with the same meaning (Dunn 2015).

[1] For an excellent introduction to the logic of character matrices, including the characteristics of 'good' characters, see Mishler (2006).

[2] These analyses are thus not based on cognate sets *per se*, since cognate sets may include forms whose meaning has shifted. Root-meaning sets exhibit some weaknesses in comparison to cognate sets as the basis of characters, e.g. they are more likely to be the result of parallel semantic shift, and they are not independent, since a presence in one set predicts absences in the other cognates of the same meaning.

2. A tree is constructed from the distance matrix, either using an explicit optimality criterion (e.g. least squares method) or a particular clustering algorithm (e.g. UPGMA, Neighbor-joining) (Felsenstein 2004: 147-175).

The main advantage of distance-based methods is that they are computationally very fast,[3] even with large numbers of languages. However, they have a serious drawback: their results do not necessarily reflect genealogical relationships, as implied by the data. The main reason for this is that the calculation of the distance matrix collapses the phylogenetically rich information of the character matrix into a small number of similarity measures between pairs of languages, with the distributional complexities of particular characters being lost. In particular, shared innovations, shared retentions, and parallel innovations are all given equal weight in calculating similarity. However, only shared innovations are evidence for subgrouping (Fox 1995:202, a.o.), making the conflation of these three sources of similarity in the calculation of the distance matrix intrinsically problematic for developing genealogical classifications.

The process of tree construction using distance-based methods thus produces a hierarchical structure in which languages are grouped according to overall similarity. Some such groups may be supported by shared innovations, but others will result from shared retentions or parallel innovations, and the different sources of support cannot be distinguished. Classifications based on distance methods thus cannot be interpreted as genealogical ones, although they may resemble accurate genealogical trees, since genealogically related languages tend to be similar. Dunn and Terrell (2012) have also found that distance-based methods are very sensitive to undetected loans.

Character-based methods, in contrast, directly employ all the information in character matrices for tree inference, without a simplifying distance matrix calculation. They examine each feature and its distribution within the dataset, optimizing its evolution on an enormous number of potential phylogenetic trees. Crucially, via the rooting of the tree, only shared innovations are used as evidence for subgrouping, meaning that they produce genealogical classifications. We briefly discuss rooting now.

The root of a phylogenetic tree represents the common ancestor of all languages included in the analysis,[4] and without a root, the direction of the flow of time on a tree (or network) is not defined, and it thus is impossible to infer genealogical subgroups. Crucially, many CP methods operate on unrooted trees, since key factors of the relevant models, such as the parsimony score and the likelihood of a tree, do not depend on the position of the root, and unrooted trees present certain computational advantages (Felsenstein, 2004). Trees (and networks) derived by either distance-based or character-based methods must thus be rooted to be interpretable historically. CP trees are usually rooted either by using a 'clock' (a model of character rates of change), or by including an outgroup in the analysis. An outgroup is one or more languages known to be outside the group being classified (the 'ingroup'), but related to the ingroup languages, and the point in the tree where

[3] To wit, lexicostatistics is a distance-based method that was feasible to implement without computers, using the percentage of shared cognates between languages as a distance metric (Swadesh, 1952).

[4] When applying the comparative method, the tree is rooted with the reconstruction of proto-sounds.

the outgroup joins the ingroup is considered *a priori* the root of the tree. The necessarily genealogically-related outgroup languages make it possible to identify the innovations that define the ingroup, and thereby identify the root. Once rooted, character states can be interpreted as innovations or retentions, with only the former supporting subgrouping.

The two most commonly used character-based methods in linguistics are parsimony and Bayesian inference. Parsimony methods infer trees by minimizing the amount of change on the tree, with different methods differing in regards to the costs associated with changes, the ordering of character states, and the weights of characters (Felsenstein 2004: 73-85). These methods do not rely on an explicit evolutionary model, unlike Bayesian methods, which we discuss next. Significantly, the Comparative Method can be viewed as a 'manual' parsimony method with a non-quantified approach to change costs and character weights. Parsimony works well when characters change sufficiently slowly and the character state space is open (i.e. there are many potential outcomes for the evolution of a given character). However, when the same states can arise repeatedly given enough time, a situation which arises when characters evolve quickly and potential character states are limited, parsimony can be misled due to parallel changes being falsely interpreted as shared innovations (Felsenstein 1978).

Bayesian Inference requires an explicit model of character evolution, as well as a set of distributions for all model parameters (i.e., prior probability distributions), representing our prior beliefs for the values of these parameters. The method then estimates the posterior probability distribution for all these parameters, in light of the data and the prior distributions. Thus, the results of a Bayesian analysis do not reply to the question "which is the phylogeny that makes these data most likely?", but rather to "what is the probability of a given phylogeny given the data and my priors?" Bayesian Inference has the advantage of being able to take into account our prior beliefs about various aspects of the model of evolution, allowing us, for example, to build into our model that certain sound changes are more common than others, without specifying exactly how much more common (as we would be forced to do with parsimony). It also gives us comparable and statistically sound estimates of our uncertainty in the results.

Finally, we address the ability of phylogenetic methods to detect and address borrowing, which results in reticulation, or network-like evolutionary structures. At this time, no methods are able to infer directly rooted phylogenetic networks, although, as with the comparative method, phylogenetic methods can be used to detect reticulation indirectly, via conflicting phylogenetic signal. It is important to point out that popular network methods (such as NeighborNet) are data visualization tools, rather than genealogical inference tools: they provide an image of how much conflicting signal there is in the data, but they cannot be interpreted historically or as a classification, since they are unrooted. The idea that one can interpret the center of the network as the root is a common misconception about these methods, as we demonstrate in §4. Even when rooted using an outgroup,[5] NeighborNets have the same drawbacks as any distance-based method.

[5] Strictly speaking, it is not possible to root trees or networks derived from distance-based methods at all, in the sense that roots of trees in historical linguistics represent the ancestral language of all languages lower in the tree. Trees developed using distance-based methods do not have roots of this type, since such trees do not represent historical or genealogical relationships.

In conclusion, there is little principled reason to select distance-based methods over character-based ones, since they ultimately require the same kinds of data (i.e. character matrices), and character-based methods are both in principle and results superior[6]. Moreover, since linguistic datasets are small in comparison to those analyzed in biosystematics, the relative speed advantage of distance-based methods is not particularly relevant, except possibly to furnish initial exploratory trees.

3. Character-based methods application to South American languages

There is a consensus in South American historical linguistics regarding the delimitation of most language families, as well as the membership of many low-level subgroups in each family (Campbell 1997). In this context, character-based methods can make a significant contribution by clarifying the relationships among low-level groups, and confirming their membership. However, since phylogenetic methods effectively presuppose the relatedness of the languages they analyze, they are not capable of identifying or evaluating long-distance relationships.

Character-based classifications of South American languages include Walker and Ribeiro's (2011) classification of Arawakan, Stark's (2018) classification of the Caribbean subgroup of Arawakan, Birchall et al.'s (2016) classification of Chapakuran, Ribeiro's (2006) of Panoan, Zariquiey et al.'s (2017) classification of the Purús Panoan subgroup, Chacon and List's (2015) classification of Tukanoan, and Michael et al.'s (2015) classification of the Tupí-Guaraní subgroup of Tupían.

Birchall et al.'s (2016) classification of Chapakuran provides a useful starting point for the discussion of these methods and the results they can yield. Based on a 207-word comparative list for ten Chapakuran varieties, the authors created 285 cognate sets, which they then subjected to a Bayesian phylogenetic analysis using BEAST, using a variety of evolutionary models, with the results summarized as a maximum clade credibility (MCC) tree. Trees were calibrated via tip-dating of data sources, and a temporal estimate for the Moré-Cojubim divergence, based on ethnohistorical information. As is evident in Fig. 1, the tree is quite well resolved, with all subgroups exhibiting greater than 0.9 posterior probability, with the exception of the Wanyam-Wari'-Oro Win subgroup.

< Figure 1 here>

Figure 1: Maximum clade credibility (MCC) tree summary of posterior probability distribution of the relaxed clock CTMC analysis of the Chapakuran family from Birchall et al. (2016).

Significantly, the authors also include a classification based on the comparative method that relies on a reconstruction of the Proto-Chapakuran segmental inventory, with subgroups defined by probative shared sound changes. Crucially, the phylogenetic classification is consistent with the sound change-based classification, although the former is considerably more detailed. Birchall et

[6] A reviewer suggests that the popularity of distance-based methods is related to their ease of use, compared to phylogenetic software, such as MrBayes and BEAST, which have a steep learning curve.

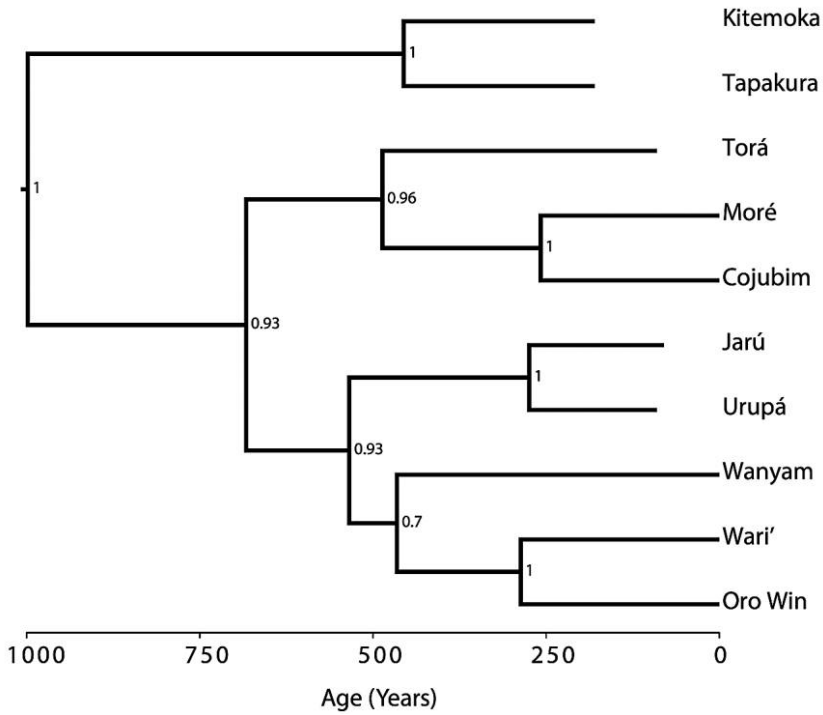

Figure 1: Maximum clade credibility (MCC) tree summary of posterior probability distribution of the relaxed clock CTMC analysis of the Chapakuran family from Birchall et al. (2016).

al.'s phylogenetic classification of Chapakuran is the state-of-the art classification of the family, with a resolution that is directly due to the use of CP methods, since the analysis of such large quantities of lexical data would not be feasible otherwise.

One way in which Birchall et al. (2016) differs from most earlier Bayesian phylogenetic analyses (e.g., Gray et al.'s (2009) classification of Austronesian) is that the characters analyzed were based on cognate sets rather than root-meaning sets. This approach was first implemented in Michael et al.'s (2015) classification of the Tupí-Guaraní, the largest subgroup of the Tupian family. This classification was based on a 543-item concept list, and was rooted with two outgroup languages, Awetí and Mawé, well-established as the most closely-related non-TG Tupían languages (Galucio et al. 2015: 231, and citations therein), yielding the tree given in Fig. 2. Crucially, the fact that this analysis was based on cognate sets, rather than root-meaning sets, meant that the sets that were the basis of character coding included all discoverable cognates, even those that had undergone semantic shift.[7]

<Figure 2 here>

Figure 2: Tupí-Guaraní Majority-rule Consensus Tree from Michael et al. (2015), with coloring corresponding to subgroups in Rodrigues and Cabral (2002)

Unlike the Chapakuran case, there is no classification of the TG languages based on the comparative method against which to evaluate the phylogenetic classification, but Rodrigues and Cabral's (2002) expert classification[8] of the TG languages into 8 low-level subgroups[9] serves as a useful validation check. As evident in the coloring in Fig. 2, Michael et al. (2015) returns most of the low-level subgroups, while, Fig. 3 compares the higher-level structure of Rodrigues and Cabral's (2002) and Michael et al.'s (2015) classifications, where in the latter case only subgroups with a posterior probability of > 0.80 have been retained as well-supported. Michael et al.'s (2015) classification provides considerably more higher-level structure, and shows that a number of higher level groups posited by Rodrigues and Cabral (2002) emerge as paraphyletic[10] in the phylogenetic classification. The significant congruence with lower-level subgroups in the expert classification encourages confidence in the phylogenetic classification, while the additional higher-level structure identifies productive directions for future research.

---

[7] While this coding scheme can be a significant improvement over the root-meaning sets in terms of character independence, it typically cannot achieve full independence, since absences in one cognate set are often inferred from the presence of a word in another cognate set that shares a meaning with words in the first cognate set.

[8] An 'expert classification' is one based on an expert's deep knowledge of a particular language family, but one that is not supported with an explicit methodology or empirical evidence.

[9] This classification is an update to Rodrigues' (1985) influential classification of the family.

[10] While members of a well-defined subgroup (or 'monophyletic group') share a most recent common ancestor that is not the ancestor of any languages outside that subgroup, the most recent common ancestor of a paraphyletic group is also the ancestor of languages *outside* the group. Often, members of a paraphyletic group are similar because of shared retentions, and can thus be misidentified as forming a monophyletic subgroup by distance-based methods, or in impressionistic classifications.

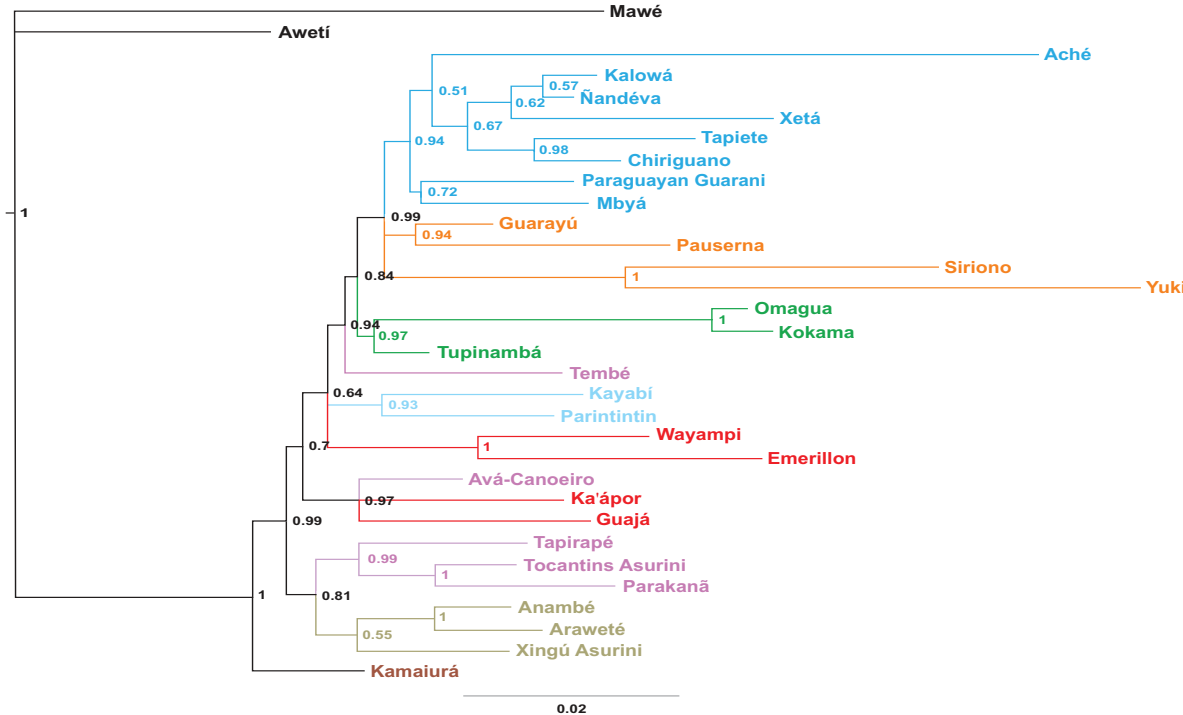

Figure 2: Tupí-Guaraní Majority-rule Consensus Tree from Michael et al. (2015), with coloring corresponding to subgroups in Rodrigues and Cabral (2002)

<Figure 3 here>

Figure 3: Comparison of higher structure in Rodrigues and Cabral's (2002) Tupí-Guaraní classification (left) with the collapsed high-confidence classification of Michael et al. (2015) (right)

Significantly, Chousou-Polydouri et al. (2016) show that the analysis of the same data on the basis of root-meaning sets produces a noticeably different classification of the TG languages, and one at odds with the expert classification, suggesting that at least with some data sets, the selection of a cognate set-based analysis over a root-meaning set-based one may be critical for obtaining reliable results.

The mention of Michael et al.'s (2015) 0.80 cutoff for considering subgroups to be well supported in their TG classification raises the question of when support may be too low for a clade to be considered significant by a phylogenetic analysis[11]. In this light, it is useful to consider Walker and Ribeiro's (2012) classification of Arawakan. Using a 100-word Swadesh list for 60 Arawakan languages, these authors carried out a Bayesian phylogenetic analysis based on root-meaning sets, relying heavily on Payne (1991) for cognacy judgments. The results of this analysis, summarized as a maximum clade credibility tree, is given in Fig. 4.

<Figure 4 here>

Figure 4: Maximum Clade Credibility tree for Arawakan from Walker and Ribeiro (2012).

Comparison of low-level subgroups in Fig. 4 with those identified in expert classifications (Aikhenvald (1999), Campbell (2012:71-77)) shows good agreement at this level, with posterior probabilities near 1. Many of the higher-level nodes, however, have less than 0.80 support; indeed, many have less than 0.50 support, meaning that most higher-level subgroups in Fig. 4 are not well supported. Once nodes with low posterior probabilities are collapsed, the classification emerges as rake-like near the root, and significantly, the lower resolution of the Arawakan tree in comparison to the Chapakuran and Tupí-Guaraní trees discussed above correlates to both a smaller concept list and a larger number of languages.[12]

[11] The significance level of any statistical test is a choice depending on the question at hand. Common significance levels used are 0.9 or 0.95 (equivalent to a p-value of 0.1 and 0.05 respectively). Michael et al. (2015) employed 0.8, as their primary goal was to put forward fruitful hypotheses for further testing.

[12] In general, the larger the number of taxa (i.e., languages), the larger the number of characters required to obtain a fully resolved tree, although this relationship is complicated by factors such as: 1) different internal branches having different number of changes; 2) different characters being informative for different time depths depending on their rate of evolution (Townsend 2007); 3) the number of character states and how they are distributed among taxa (Bordewich et al. 2018), and 4) the presence of characters with conflicting phylogenetic signal. Of course, if the tree in fact exhibits rake-like organization ('hard polytomies'), no number of characters will resolve them.

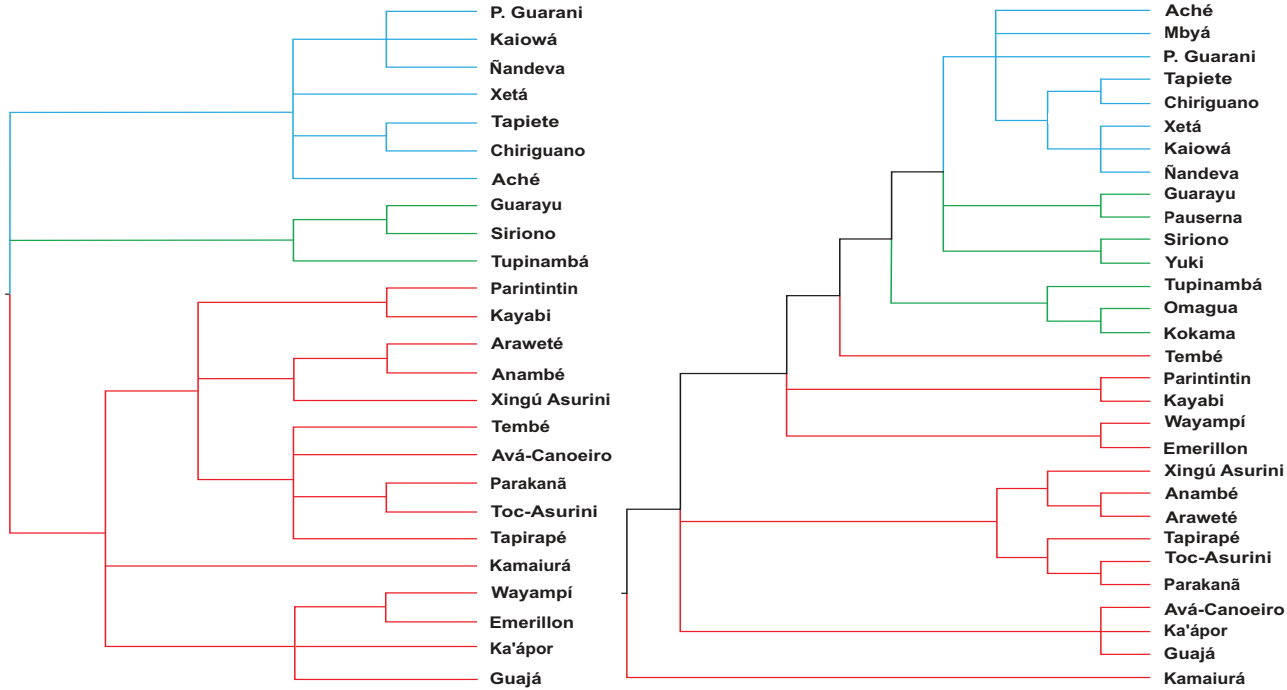

Figure 3: Comparison of higher structure in Rodrigues and Cabral's (2002) Tupí-Guaraní classification (left) with the collapsed high-confidence classification of Michael et al. (2015) (right)

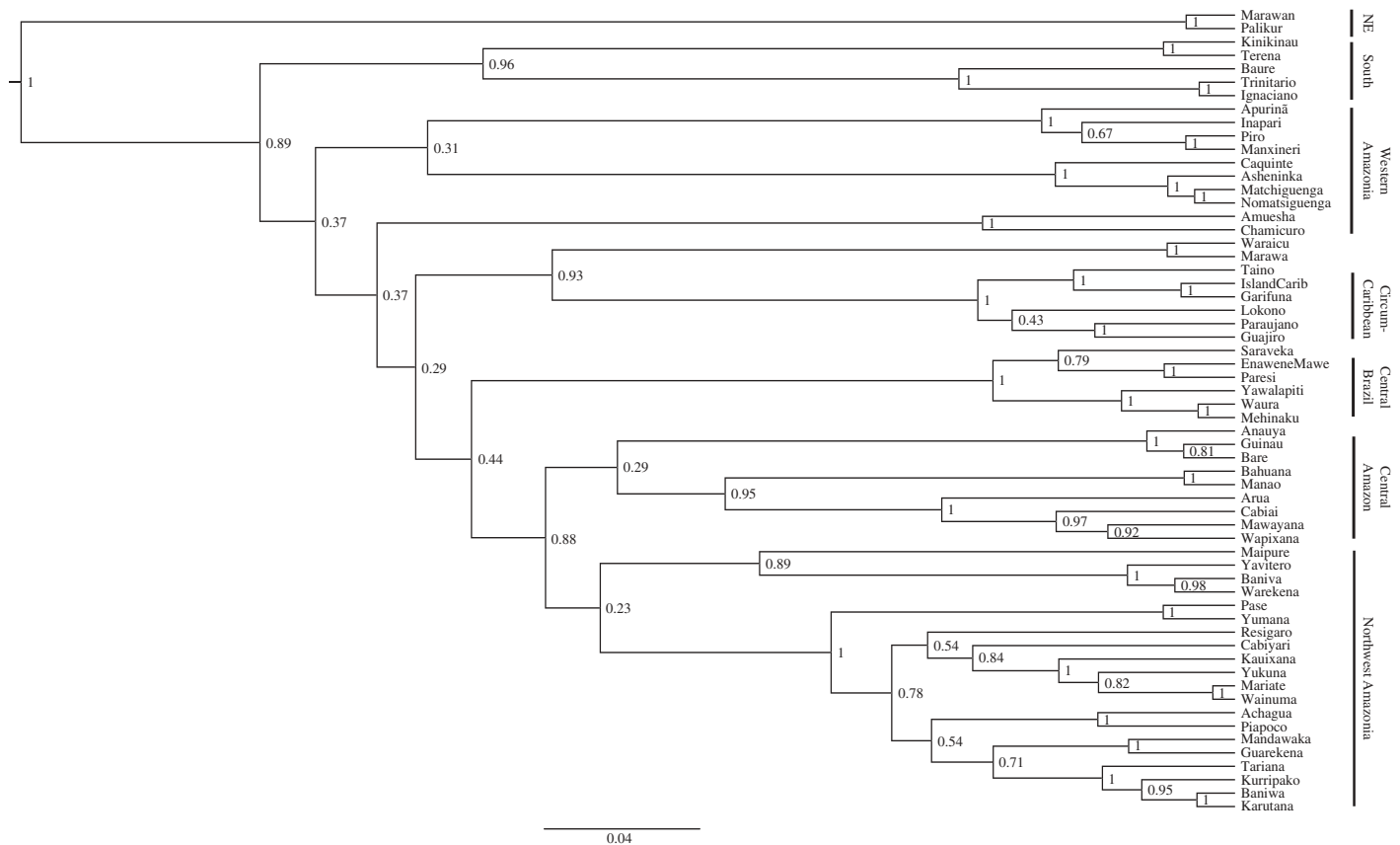

Figure 4: Maximum Clade Credibility tree for Arawakan from Walker and Ribeiro (2012).

In this light, it is instructive that Stark's (2018) Bayesian analysis of the Caribbean sub-group, based on a 736-item comparative list, yields a well-supported internal classification that provides strong evidence for a Lokono-Wayuu-Añun subgroup, which is not well supported in Walker and Ribeiro (2012). This classification, which employed Achagua, Baniwa, Palikur, and Wapishana as outgroup languages is given in Fig. 5. Stark's markedly higher posterior probabilities are presumably due to her more informative comparative list.

<Figure 5 here>

Figure 5: Maximum clade credibility tree for Caribbean Arawak from Stark (2018)

Despite the relative dearth of well-supported high-level nodes, Walker and Ribeiro's (2012) classification includes some intriguing features, especially at mid-level nodes. For example, the Waraiku-Marawa subgroup is found to form a clade with the well-established Caribbean subgroup, while the Manao-Bahuana subgroup is identified as forming a clade with a subgroup that consists of Wapishana and Mawayana (itself an uncontroversial grouping), but also Cariay and Aruã. Among the southern Arawakan languages, Yanesha' and Chamicuro are identified as forming a low-level group, while Saraveka and Paresi form a clade with the Xinguan Arawakan languages. All but the last of these subgroups were hitherto little-suspected, and merit further investigation. Walker and Ribeiro's analysis also provides reasonably strong support for a large subgroup consisting of the bulk of the northern Arawakan languages, excluding the members of the extended Caribbean group (see above). A similar subgrouping proposal is found in Aikhenvald's expert classification (but including Manao and Bahuana, which are excluded from this subgroup in the phylogenetic analysis). Walker and Ribeiro also identify the Marawan-Palikur subgroup as sister to the remainder of the family, which constitutes a well-defined subgroup.

We now turn to phylogenetic classifications of Panoan, which includes the first published phylogenetic analysis of a South American language, Ribeiro (2006). This paper presents a Bayesian analysis based on the 100-item Swadesh concept list, which yields a large rake, appearing to resolve only a few subgroups, such as the Headwaters group, a clade consisting of Matsés and its close relatives, and one consisting of Shipibo and its neighbors. As Ribeiro (ibid., 177) himself notes, the tree is unrooted,[13] and as such, no classification can be inferred from it (see §2), although the author does present one. Regardless of the rooting issue, we see, however, that a 100-item concept list fails to provide a well-resolved tree for the family. Zariquiey (2017) presents a character-based Bayesian phylogenetic classification of nine Panoan varieties of the Purus River region, based on a 180-item Swadesh list plus 68 binary typological characters. Lexical characters yield a relatively rake-like structure, reflecting in part the considerable lexical similarity among the varieties.

We conclude with a discussion of Chacon and List's (2015) Tukanoan classification, which uses parsimony methods in conjunction with classic phonological reconstruction, with the characters

[13] The tree most probably exhibits the default setting of the MrBayes application, which is to consider the first language in the input file, Amawaka, as an outgroup. This has no linguistic justification.

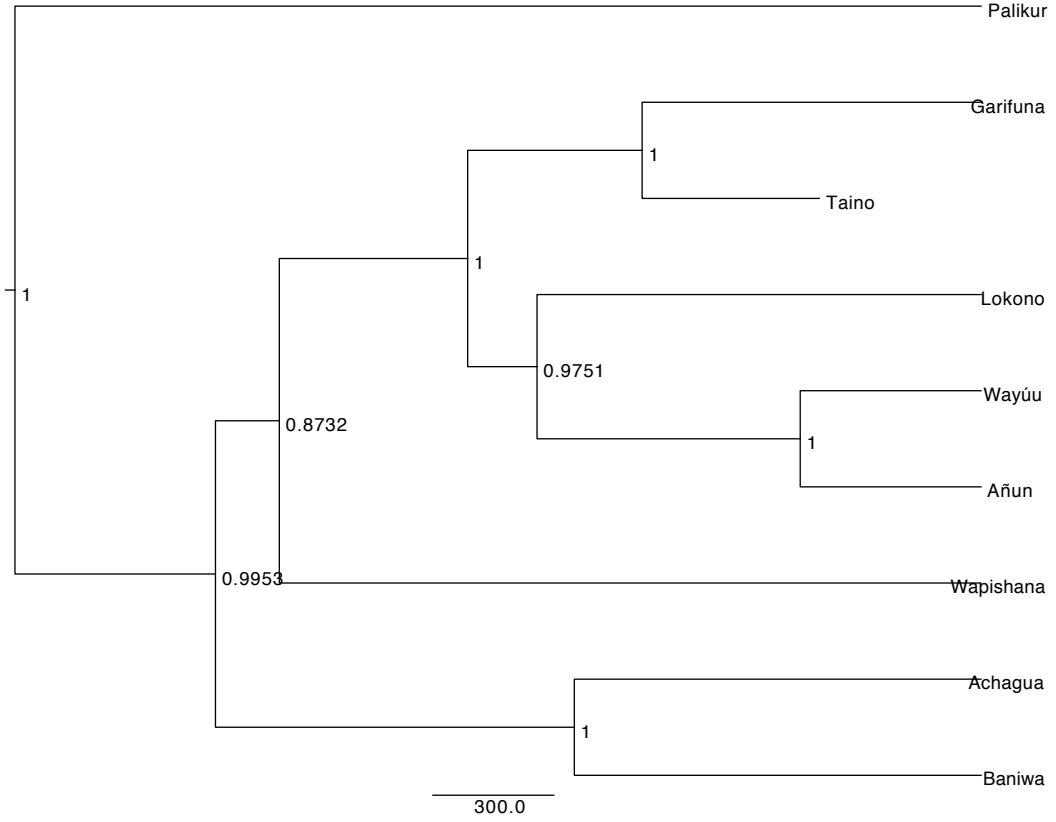

Figure 5: Maximum clade credibility tree for Caribbean Arawak from Stark (2018)

analyzed being multi-state correspondence set (i.e. proto-sound reflex) characters. Building on Chacon's (2014) reconstruction of the Proto-Tukanoan (PT) consonantal inventory, this work identifies the set of sound changes relating the segments of PT to their reflexes in modern Tukanoan languages, and evaluates the efficacy of a number of parsimony-based methods for their resolving power and their ability to successfully reconstruct the sound changes on candidate trees. The most successful of these is one that uses a directed weighted state transition (DiWeST) network linking possible sound changes to infer a tree that minimizes the number of independent phonological innovations, given in Fig. 6. This tree differs in a number of notable ways from Chacon's (2014) classification, which is likewise based on shared phonological innovations, but was inferred 'manually', including identifying Kubeo and Tanimuka as forming a first-branching subgroup of Eastern Tukanoan, and grouping Máihɨ̃ki with Koreguahe and Kueretu. Chacon (p.c.) considers the 'manual' consensus tree given in Chacon and List (2015: 198) based on the DiWeST and Chacon (2014), to be the state-of-the-art classification of the family.

<Figure 6 here>

Figure 6: Parsimony-based classification of Tukanoan using a directed weighted state transition network (DiWeST) from Chacon and List (2015)

4. Distance-based methods application to South American languages

Because of the weaknesses of distance-based methods discussed in §2, genealogical classifications derived from them should be approached with caution. In those cases where distance-based and character-based classifications for South American language families exist, they differ, with the former regularly diverging from expert classifications and classifications based on the comparative method.

Differences between distance- and character-based analyses are illustrated by a Neighbor-Joining (NJ) tree based on the same TG data analyzed in Michael et al. (2015).[14] Comparing the NJ tree given in Fig. 7 with the Bayesian tree based on the same data (Fig. 2), we see that while there are clades in common between them (some of them marked in blue), there are also important differences (some of them marked in red). For example, Kamaiura is sister to all other Tupí-Guaraní languages in the Bayesian tree, while in the NJ tree it is sister to Ava Canoeiro. Another important difference is the position of Tupinambá, which is very closely related to Omagua and Kokama (Michael and O'Hagan 2016), but does not not form a clade with them in the NJ tree, although in both expert and the Bayesian classifications it does. This is related to the fact that in the NJ tree, all the languages that have undergone extensive lexical change (Ache, Siriono, Yuki, Xeta, Omagua, and Kokama) are the first to diverge, leaving a clade of all the remaining TG languages (presumably linked by shared retentions).

<Figure 7>

[14] The Neighbor-Joining tree was built with SplitsTree 4 (Huson & Bryant, 2006)), based on uncorrected distances calculated from the character matrix used in Michael et al. (2015). It was rooted with Mawe in FigTree v.1.4.4 (Rambaut, 2018).

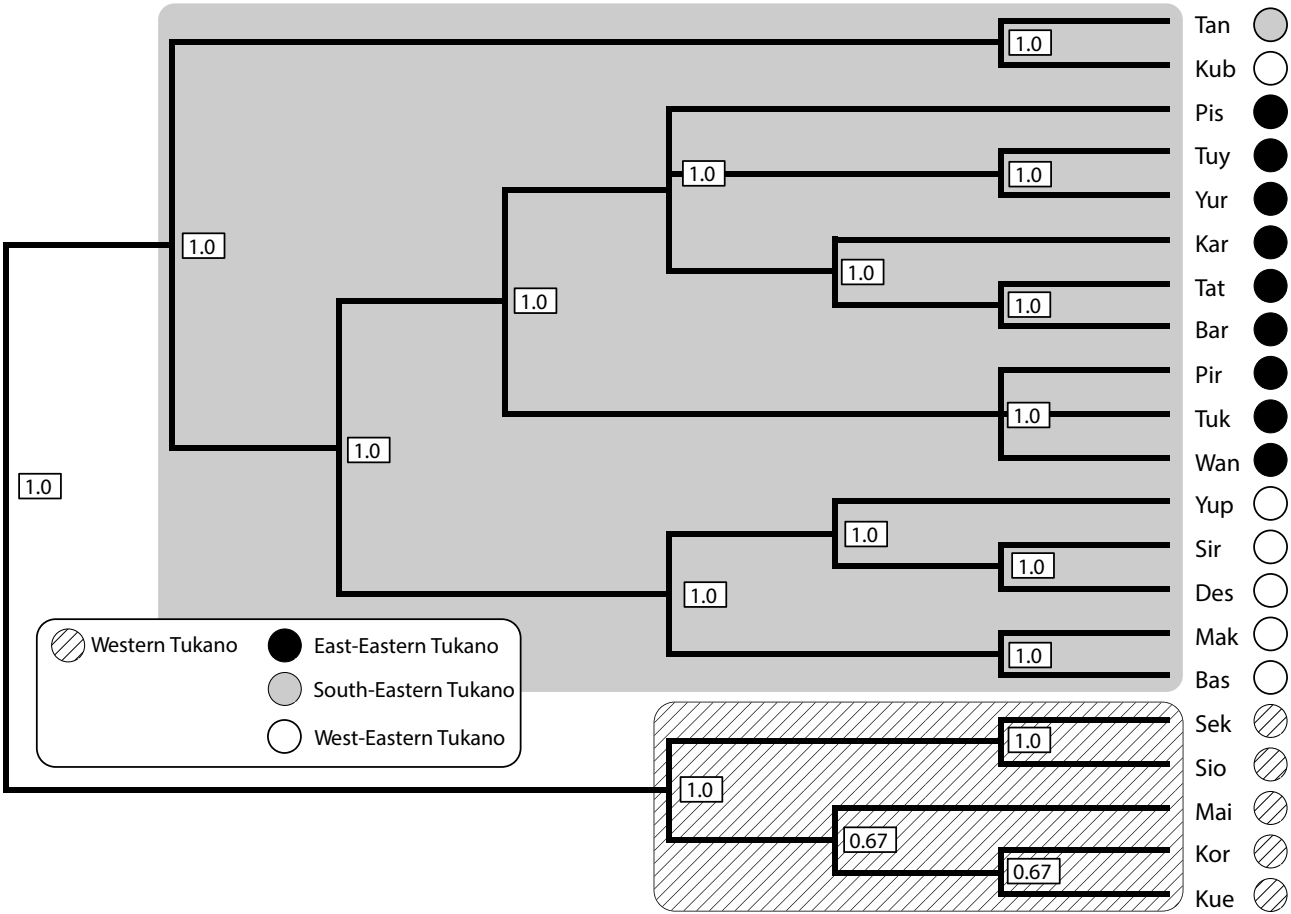

Figure 6: Parsimony-based classification of Tukanoan using a directed weighted state transition network (DiWeST) from Chacon and List (2015)

Figure 7: A neighbor-joined distance-based classification of Tupí-Guaraní

A similar divergence between distance- and character-based classifications is exemplified in competing classifications of the Tupian family. Walker et al. (2012) provides a distance-based classification of the Tupían family yielded by the Automated Similarity Judgment Program project (Holman et al. 2011), given in Fig. 8. This project produces distance-based classifications based on Levenshtein distances between words from a 40-item concept list for a majority of the world's languages, where Levenshtein distances between words correspond to the number of symbol substitutions necessary to convert one of two words being compared to the other. Words used in the ASJP analysis are represented in 'ASJPcode', a reduced inventory of symbols by which all IPA symbols are binned into 41 symbols found on a typical QWERTY keyboard. Distances between languages are thus not based on percentages of shared cognates, but on a coarse-grained measure of phonological similarity of words with the same meaning. This method thus systematically fails to distinguish between cognacy and accidental phonological similarity.

<Figure 8>

Figure 8: ASJP classification of the Tupian family from Walker et al. (2012)

Focusing first on the Tupí-Guaraní sub-group, for which both expert and phylogenetic classifications exist, we see that the ASJP classification does identify some recognized low-level groupings, but that the overall structure of the tree diverges wildly from these other two types of classifications (see §3). We focus on some examples here to illustrate the nature of the divergences involved.

We see a similar issue with the relative positions of Tupinambá, Omagua, and Kukama as that found in the neighbor-joining classification given in Fig. 8, but likely compounded by the fact that the language contact situation that gave rise to Proto-Omagua-Kukama involved significant changes in the phonological form of lexical items and the appearance of frozen morphology, resulting in cognate forms between Tupinambá on the one hand, and Omagua and Kukama on the other, being phonologically relatively dissimilar in comparison with other sets of closely-related TG languages. This would lead to large -- and for purposes of classification, misleading -- Levenshtein distances between the languages. Strikingly, Tupinambá is classified with Parintintin, which is not supported either by expert classifications or character-based analyses.

The converse problem is likely evident in the classification of Kayabí and Kamayurá as forming a low-level clade, a proposal likewise unsupported by phylogenetic or expert classifications. In this case, the classification is probably due to the fact that these are two of the modest number of TG languages that lack a prenasalized stop series (e.g. mb, nd) and instead exhibit the corresponding nasal stops in cognate forms, which would lead to smaller Levenshtein distances between these languages for words that exhibit prenasalized stops in most other languages of the family. There is no reason to believe that the absence of the pre-nasalized stops in these languages is a shared innovation, however, and both character-based and expert classifications agree in grouping Kayabí with Parintinin, and not with Kamayurá.

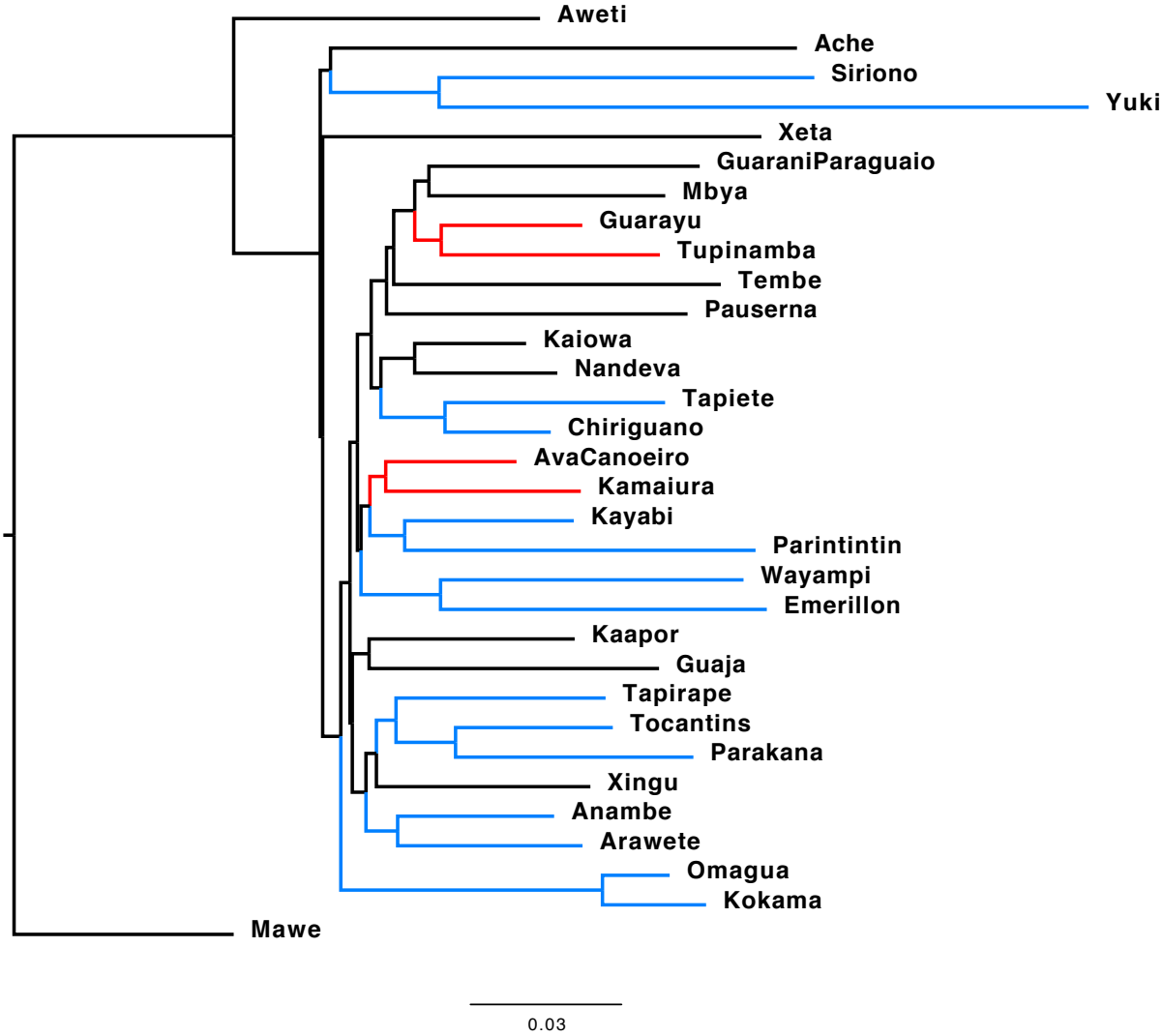

Figure 7: A neighbor-joined distance-based classification of Tupí-Guaraní

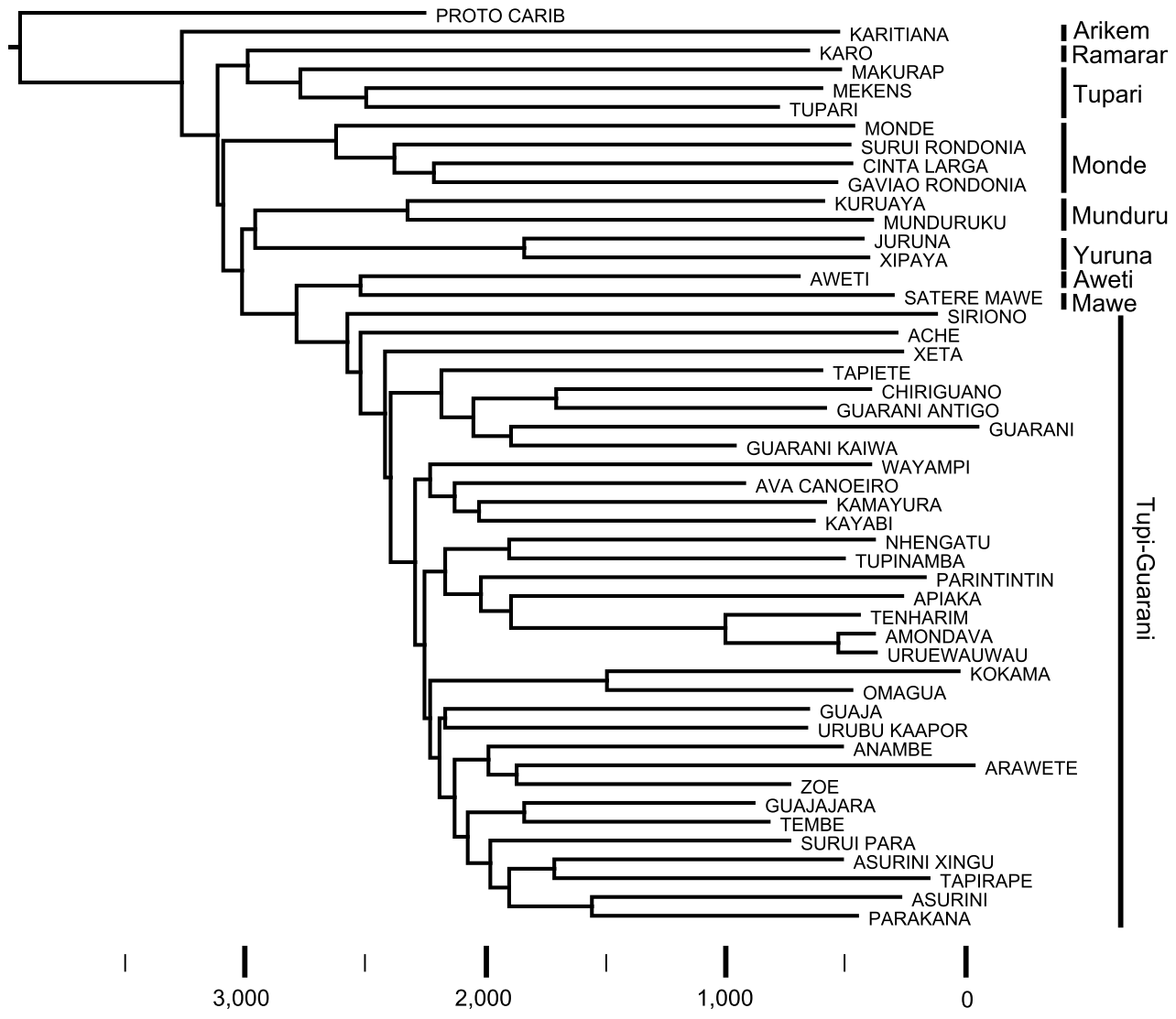

Figure 8: ASJP classification of the Tupian family from Walker et al. (2012)

It is also instructive to compare the ASJP classification with Galucio et al.'s (2015) classification of Tupían, a distance-based classification that relies on a 100-item concept list and cognacy judgments of Tupían experts. Galucio et al. coincides with phylogenetic (Fig. 2) and expert classifications, given in Fig. 9, in classifying Awetí as sister TG, and Mawé forming a sister to the Awetí-TG group. The ASJP classification, however, finds an Mawé-Awetí clade that is sister to TG. The ASJP classification also identifies Karo as forming a clade with the Tuparí group, and the Juruna-Xipaya and Mundurukú-Kuruaya clades as forming a subgroup, none of which is not supported by either the expert classification or Galucio et al. (2015).

<Figure 9>

Figure 9: Expert classification of the Tupí family from Galucio et al. (2015)

Thus, while both the ASJP classification and the Galucio et al. (2015) classifications are distance-based classifications, and suffer from the weaknesses inherent to such methods, the latter, based on a longer concept list and expert cognacy judgments, produces a classification more in line with expert classifications and character-based classifications than the ASJP method. Note also that Walker et al. (2012) roots Tupian using Cariban as an outgroup. The justification for this is weak, since a putative Tupian-Cariban relationship has yet to be demonstrated to the satisfaction of specialists (Campbell 1997: 201-202, Michael to appear).[15]

5. Conclusion

Phylogenetic methods are a recent addition to the toolkit of South Americanist linguists, but they are already yielding important advances in our understanding of the internal classification of families of the continent. While lexical data has been the principal empirical focus of phylogenetic analyses to this point, we see preliminary efforts to extend these methods to sound change (Chacon and List 2015), and morphosyntactic data (Chousou-Polydouri et al. 2016), which promise to complement lexical data in useful ways.

For linguists interested in applying these methods, it is important to keep certain critical points in mind. First, distance-based methods are limited in their ability to contribute to our understanding of genealogical relationships, principally due to the fact that they do not yield subgroups based on shared innovations, but rather, on a single overall similarity measure. Character-based methods, in contrast, yield classifications based on shared innovations, and are thus compatible with this key insight of the comparative method.

Second, the productive implementation of character-based methods requires selecting character sets that are adequate for producing well-resolved and accurate trees. Root-meaning sets based on small comparative lists may not be adequate for this purpose, especially for larger language

[15] Strictly speaking, only trees developed using character-based methods have roots in the historical/evolutionary sense (see §2, and especially fn. 5), and with such methods, outgroups must be genealogically related to the ingroup languages in order to be able to the root them. To the degree that outgroups are intended to root trees developed using distance-based methods, the same requirement would apply.

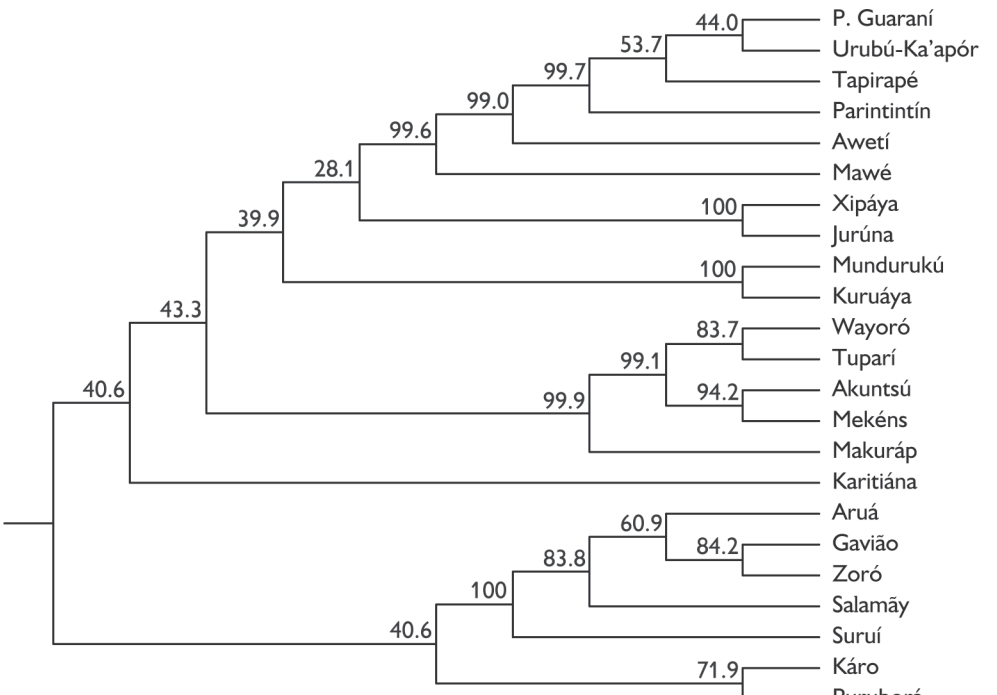

Figure 9: Expert classification of the Tupí family from Galucio et al. (2015)

families. This suggests that an important goal for South Americanist linguists using phylogenetic methods is the assembly of sufficiently large comparative lexical data sets and the corresponding cognate sets. This is a highly labor-intensive task, which encourages collaboration among specialists on the relevant language families.

And third, care must be taken in how phylogenetic trees are rooted. For example, while the use of outgroups is in many ways ideal, the use of highly speculative long-distance relationships to justify using one major language family to root the phylogenetic tree of another (e.g., treating Cariban as an outgroup for rooting Tupian) is highly dubious. In cases such as these, alternate rooting techniques, such as including a clock or, character polarization based on the directionality of sound change, should be employed.

For consumers of results from the application of these methods, it is similarly important to keep the above points in mind as critical readers. Do particular publications make use of reliable methods? Are the analytical choices justified and reasonable? Are the datasets adequate for the analytical purposes at hand? Almost any phylogenetic method and any comparative dataset is capable of producing a classification, but whether that classification is trustworthy depends on how it was obtained.

Finally, we wish to observe that the successful application of phylogenetic methods depends on two critical, but sometimes undervalued, aspects of linguistic research: 1) the development of high-quality linguistic documentation, especially lexical documentation; and 2) long-term collaboration between linguists and indigenous communities. Progress in our understanding of the classification of South American languages thus depends on substantive support for the necessary collaborations across the many communities of the continent, and support for lexical documentation in particular as a valued aspect of basic linguistic research.

6. References

Birchall, J., Dunn M., & Greenhill, S. (2016). A combined comparative and phylogenetic analysis of the Chapacuran language family. *International Journal of American Linguistics, 82(3)*, 255-284.

Campbell, L. (1997). *American Indian languages: the historical linguistics of Native America*. Oxford University Press.

Chacon, T., & List, J.M. (2015). Improved computational models of sound change shed light on the history of the Tukanoan languages. *Journal of Language Relationships, 13(3)*, 177-203.

Chang, W., Cathcart, C., Hall, D., & Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, *91*(1), 194-244.

Chousou-Polydouri, N., Gasparini, N., Michael, L., O'Hagan, Z., Rose, F. (2016). Reconstructing negation morphemes and constructions in Tupí-Guaraní. Talk presented at Amazónicas VI, Letícia, Colombia.

DeSalle, R. (2006). What's in a character? *Journal of Biomedical Informatics*, *39*(1), 6–17.

Drummond, A. J., & Bouckaert, R. R. (2015). *Bayesian evolutionary analysis with BEAST*. Cambridge University Press.

Dunn, M. (2015). Language phylogenies. In Bowern, C., & Evans, B. (Eds.), *The Routledge handbook of historical linguistics*. Routledge, pp 208-229.

Dunn, M., & Terrill, A. (2012). Assessing the lexical evidence for a Central Solomons Papuan family using the Oswalt Monte Carlo Test. *Diachronica*, *29*(1), 1-27.

Epps, P. (2009). Language classification, language contact, and Amazonian prehistory. *Language and Linguistics Compass, 3(2)*, 581-606.

Epps, P., & Michael, L. (2017). The areal linguistics of Amazonia. In Hickey, R. (Ed.), *The Cambridge Handbook of Areal Linguistics*. Cambridge University Press, pp. 934-963.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic zoology*, *27*(4), 401-410.

Felsenstein, J. (2004). Inferring phylogenies. Sunderland, Massachusetts: Sinauer Associates, Inc.

Fox, A. (1995). *Linguistic reconstruction: an introduction to theory and method*. Oxford University Press on Demand.

Galucio, A. V., Meira, S., Birchall, J., Moore, D., Gabas Júnior, N., Drude, S., Storto, S., Picanço, G., & Rodrigues, C. R. (2015). Genealogical relations and lexical distances within the Tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, *10*(2), 229-274.

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., , Sauppe, S., Jung, H., Bakker, D., Brown, P., & Belyaev, O. (2011). Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, *52*(6), 841-875.

Huson, D. H., & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, *23*(2), 254–267.

Michael, L. (to appear). The classification of South American languages: The state of the art. *Annual Review of Linguistics*.

Michael, L., & O'Hagan, Z. (2016). A linguistic analysis of Old Omagua ecclesiastical texts. *Cadernos do Etnolingüística: Serie Monografías*.

Michael, L., Chousou-Polydouri., N, Bartholomei, K., Donnelly, E., Wauters, V., Meira, S. & O'Hagan, Z. (2015). A Bayesian Phylogenetic Classification of Tupí-Guaraní. LIAMES 15(2): 193-221.

Mishler, B. D. (2006). The logic of the data matrix in phylogenetic analysis. In V. A. Albert (Ed.), *Parsimony, Phylogeny, and Genomics*. Oxford University Press, pp. 57-70.

Nichols, J., & Warnow, T. (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass, 2(5)*, pp. 760-820.

Payne, D. L. (1991). A classification of Maipuran (Arawakan) languages based on shared lexical. In Derbyshire, D., & Pullum, G. (Eds.), *Handbook of Amazonian languages (Vol. 3)*. Walter de Gruyter, 355-499.

Rodrigues, A. D. (1984). Relações internas na família lingüística Tupí-Guaraní. *Revista de antropologia*, 33-53.

Rambaut, A. (2018). *FigTree v.1.4.4*. Retrieved from https://github.com/rambaut/figtree/releases

Ribeiro, L. A. A. (2006). Uma proposta de classificação interna das línguas da família Pano. *Revista Investigações, 19*(2), pp. 157-188.

Stark, T. E. (2018). Caribbean Northern Arawak Person Marking and Alignment: a Comparative and Diachronic Analysis. (Unpublished doctoral dissertation). UC Berkeley, Berkeley, USA.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, *96*(4), 452–463.

Walker, R. S., & Ribeiro, L. A. (2011). Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society of London B: Biological Sciences*, *278*(1718), 2562-2567.

Walker, R. S., Wichmann, S., Mailund, T., & Atkisson, C. J. (2012). Cultural phylogenetics of the Tupi language family in lowland South America. *PloS one*, *7*(4), 1-9.

Wheeler, T.J. 2009. Large-scale neighbor-joining with NINJA. In S.L. Salzberg and T. Warnow (Eds.), Proceedings of the 9th Workshop on Algorithms in Bioinformatics. WABI 2009, pp. 375-389. Springer, Berlin.

Zariquiey, R., Vásquez, A., & Tello, G. (2017). Lenguas y dialectos pano del Purús: una aproximación filogenética. *Lexis*, *41*(1), 83-120.